DOCUMENT RESUME

ED 334 229 TM 016 737

AUTHOR Ryan, Katherine E.
TITLE The Performance of the Mantel-Haenszel Procedure.
PUB DATE Apr 90
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Black Students; Comparative Analysis; Context Effect; Correlation; *Estimation (Mathematics); Grade 8; *Item Bias; Junior High Schools; *Junior High School Students; *Mathematics Tests; Reliability; Robustness (Statistics); *Sample Size; *Standardized Tests; Test Items; White Students
IDENTIFIERS *Mantel Haenszel Procedure; Second International Mathematics Study

ABSTRACT

In an investigation of item bias, the stability of Mantel-Haenszel (MH) estimates across different samples of test takers and different sample sizes and the robustness of the MH procedure with respect to item context effects were investigated. Data from the Second International Mathematics Study (1985) were analyzed. The data consisted of responses to 40 core items on the mathematics tests and one of four 35-item rotated mathematics tests for a core sample of 670 black and 5,015 white eighth graders in the United States. Most analyses were performed with a sample of just over 100 blacks and slightly over 1,000 whites. Correlations between different samples of test takers were low, suggesting that relatively larger sample sizes were necessary for stable estimates. Correlational analyses also suggest that the MH procedure was robust to item context effects. Four of the 40 core items were identified as functioning differently for black and white test takers. The MH procedure is useful in detecting differential item functioning; however, further investigation is needed to determine the sample size necessary for stable estimates. Seven tables present study results. A 21-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334229

"THE PERFORMANCE OF THE MANTEL-HAENSZEL PROCEDURE"

Katherine E. Ryan

Metritech, Inc.

111 North Market Street, Champaign, Il 61820

To be presented to the 1990 AERA Annual Meetings

Session Title: Differential Item Functioning

Boston, 16 April, 1990

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KATHERINE E. RYAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM016737

BEST COPY AVAILABLE

INTRODUCTION

Bias in testing has been a major issue in educational measurement (Cole, 1981). One aspect of this issue is item bias. Numerous methods have been proposed for detecting test items that do not function in the same manner for specific subgroups (Shepard, Camilli, & Williams, 1985; Linn, Levine, Hastings & Wardrop, 1981; Camilli, 1979; Scheuneman, 1979; Lord, 1977; Angloff & Ford, 1973; Green & Draper, 1972). However, as the performance of particular ethnic and racial subgroups continues to be consistently lower on standardized achievement tests (Linn & Drasgow, 1987; Wigdor & Garner, 1982), "item bias" continues to be controversial and confusing. Recent literature (Holland & Thayer, 1986) suggests that even the traditional terminology such as item bias contributes to this problem. The term "differential item functioning" (DIF) suggested by Holland and Thayer (1986) will be used rather than "item bias" to describe the analyses conducted in this study.

The Mantel-Haenszel (MH) procedure has recently been suggested as an alternative procedure to IRT methods in investigating differential item functioning (Holland & Thayer, 1986; McPeek & Wild, 1986). The MH procedure was originally devised for retrospective epidemiological studies to investigate the relationship between the presence or absence of a potential risk factor and the occurrence disease (Mantel & Haenszel, 1959). The procedure has been used to detect differential item functioning on tests of

educational achievement (Zwick & Ericikan, 1989; McPeek & Wild, 1987).

However, there are few studies examining the performance of the Mantel-Haenszel procedure. McPeek and Wild (1986) investigated various flagging criteria for the MH procedure, the characteristics of flagged items, and the stability of the MH estimates with different test takers. When examining differential item functioning among black and white test takers, they found only 57% of the items were flagged in both samples and, in particular, verbal items replicated better than quantitative items. Zwick and Ericikan (1989) found that the use of more rigorous matching criteria did not reduce the number of DIF items and the MH chi-square statistics differed across the analyses using these different criteria. No studies have examined whether the Mantel-Haenszel estimates are robust to item context effects.

The purpose of this study is twofold: (a) to examine the stability of the MH estimates across different samples of test takers as well as across different sample sizes; and (b) to investigate whether the MH procedure is robust with respect to item context effects.

METHODS

Sample

Part of the data collected from the Second International Mathematics Study (SIMS) (1985) were analyzed. The data consisted of responses to the 40 item core

mathematics test (designated as Core) and the four 35 item rotated mathematics tests completed by the black and white eighth grade U.S.students (670 black students and 5,015 white students) in the sample. Each of the four rotated tests (designated as TF1S, TF2S, TF3S, or TF4S), were randomly assigned to approximately one quarter of the sample. Thus, each examinee completed a 75 item test (designated as TF1L, TF2L, TF3L, or TF4L) consisting of the 40 core test and one of the four 35 item rotated tests.

The item pool consisted of 64 arithmetic items, 42 algebra items, 42 geometry items, 26 measurement items, and 18 statistics items. Items were also classified by Bloom's (1956) behavioral levels. The tests were designed to be parallel in content and difficulty. Eight items from each content area (with the exception of statistics) were assigned to the core by stratified random assignment. The rest of the items were randomly assigned to each of the 35 item rotated forms. A summary of the descriptive statistics for each of the forms is reported in Table 1.

Insert Table 1 here

Mantel-Haenszel Procedure

To examine differential item functioning, the Mantel-Haenszel procedure and other indices proposed by Holland (1985) were calculated for all forms of the test. The Mantel-Haenszel procedure consists of the Mantel-Haenszel common-odds-ratio (MHODDS), and the Mantel-Haenszel chi-

square statistic (MHCHIX). Holland (1985) suggests transforming the MHODDS to the ETS delta metric (MH D-DIF) and to calculate the standard error of the common-odds ratio (Phillips & Holland, 1987). These indices are calculated for all items of interest and used to evaluate item functioning. Calculations of all these indices have been describe in detail elsewhere (See Holland, 1985; Holland & Thayer,1986; and Zwick & Ericikan, 1989).

Detecting DIF with the MH procedure is based on the notion of comparing item functioning for only comparable group members. Examinees are classified as either focal group members (subgroup of interest) or reference group members (standard to compare performances of the focal group) on the basis of group membership. Test takers are then matched on relevant criteria, such as total test score, instructional history, etc. before comparison of performance on the item. When matching individuals on the basis of total test scores, there may be as many score groups as there are possible scores on the test. In other words, on a 40 item test, there are potentially 41 score groups (0 to 40).

The MHODDS estimate is interpreted as the average amount by which the odds that a reference group member is correct on an item is larger than the odds for a comparable member of the focal group (Holland & Thayer, 1986). Items with a MH D-DIF equal to zero are of equal difficulty for matched groups of test-takers. A negative value for MH D-DIF indicates that the item is easier for the reference group

than for the focal group. The alternative interpretation is appropriate for positive MH D-DIF values.

The MHCHIX has an approximate chi-square distribution with one degree of freedom under the Ho: of no differential item functioning on item j for matched reference and focal group members. This statistic will identify differential item functioning that favors either subgroup when the MHODDS is significantly different than 1.0.

Mantel-Haenszel Analyses

A computer program written in Fortran (Control Data Systems,1985) was used to calculate the MHCHIX, MHODDS, the MH D-DIFF, and the standard error of the common-odds-ratio indices for all the Mantel-Haenszel analyses conducted. The item under study wasincluded in the matching criterion. Any item with a MHCHIX > 3.84 ($p < .05$ for a chi-square with 1 degree of freedom) and an MHODDS > 1.6 or <.625 was flagged. Table 2 provides a complete description of all the analyses including test form, acronym, sample sizes, matching criterion, and items tested in each analysis.

Insert Table 2 here

To provide a baseline, white-white comparisons (Set 1) were conducted with sample sizes similar to the black-white comparisons for the core items. A random sample of 670 white examinees was selected as the focal group while the remaining white test takers were the reference group for the large sample MH analysis with the core test items. In

addition, four small sample analyses with the 40 core test items were conducted with each quarter of the sample that completed the same rotated test form. (The four focal groups were from the random sample of 670 white test takers). For all five of these analyses, the criterion was total test score on the 40 core test items (Criterion 1).

The design of the analyses in Sets 1 and 2 are parallel, except that a smaller group of whites served as the focal group in Set 1 and blacks were designated as the focal group in Set 2. To examine sample size effect, five MH analyses were conducted for Set 2. The large sample (S) MH analysis tested the 40 core test items. The design was based on group membership (race) with the total score on the 40 item core test (Criterion 1) designated as the matching criterion. To examine the MH estimates from different samples of test takers taking the same items, the same classification and criterion testing the 40 core test items were used in conducting four separate analyses. Each of the small samples (s1, s2, s3, s4) completed the same rotated test form.

Further, to examine item context effects, four separate analyses (Set 3) were conducted with the small samples (s1, s2, s3, and s4), each analysis examining one test of 75 items. The four 75 item tests were comprised of the common core 40 items and one of the rotated test forms, TF1, TF2, TF3, or TF4, respectively. The test takers were, as in Set 2, classified on the basis of group membership (white or

black) while total test score on the respective 75 item test served as the criterion (Criterion 2).

In addition, to replicate item context effects, four separate analyses (Set 4) were conducted with each quarter of the sample (s1,s2,s3, and s4), with each analysis involving one of the 35 item rotated test forms. Here, examinees were classified on the basis of group membership (black or white) and total score on the respective 35 item test was designated as the matching criterion (Criterion 3).

Analyses of MH Indices

To examine the performance of the MH indices, two basic designs were used. The analyses focus on 1) the MH D-DIF or MHCHIX from the 40 common core items, or 2) the MH D-DIF or MHCHIX from the 35 item rotated test forms.

The means and standard deviations of the MHCHIX and the MH D-DIF values for the forty common core items were calculated for Sets 1, 2, and 3 and for the corresponding thirty-five item rotated forms from Sets 3 and 4.

Pearson product-moment and Spearman rank-order correlations were computed to investigate the stability of the MHCHIX and MH D-DIF indices, respectively from selected MH analyses.

For Set 1, the white-white comparisons, the correlations among the MH D-DIF and the MHCHIX from the five analyses were calculated to provide a baseline. The correlations were based on the indices from the forty common core items.

For Sets 2 and 3, the correlations among the MH D-DIF and MHCHIX from the nine analyses were computed to examine item context and sample size effects. The correlations were based on the indices from the forty common core items. From Set 3 with the four seventy-five items tests, only the MH D-DIF and MHCHIX values from the 40 common core items were used in calculating the correlations. In other words, the indices from the rotated test forms were deleted from the this part of the correlational analyses.

Lastly, the correlations among the MH D-DIF and MHCHIX for the corresponding rotated tests from Sets 3 and 4 were calculated to replicate item context effects. The correlations were based on the indices from corresponding thirty-five item rotated test forms. For example, the MH D-DIF and MHCHIX values for TF1S and TF1L, TF2S and TF2L, etc. were correlated.

RESULTS

Descriptive Statistics For the MH D-DIF and MHCHIX Values

Table 3 lists the means and standard deviations for the MH D-DIF and the MHCHIX statistics for all sets of analyses for the first 40 items.

Insert Table 3 here

As can be seen in Table 3, while the means for MH D-DIF values are close to 0 for Set 1 (white-white comparisons), they are larger for Sets 2 and 3. The matching criterion (total test score on common core items, Criterion 1) for

Sets 1 and 2 was identical. The means are largest for Set 3 where the MH D-DIF and MHCHIX values were calculated within 75 items (each form with a unique set of 35 items) and the matching criterion was total test score on the 75 items under study (Criterion 2). The standard deviations of the MH D-DIF values for the black-white comparisons (Sets 2,and 3) are all larger than the corresponding white-white comparisons (Set 1). For the MHCHIX statistics, again the means and standard deviations for the black-white comparisons are larger than the white-white comparisons.

Table 4 contains the means and standard deviations for the MH D-DIF and the MHCHIX statistics for the corresponding rotated test forms. The MH D-DIF and MHCHIX values for all "S" forms were calculated with total test score on the rotated test form items under study as the criterion (Criterion 3). The MH D-DIF and MHCHIX values for all "L" forms were calculated with the total test score on the 75 items under study (Criterion 2) as the matching criterion.

Insert Table 4 here

As shown in Table 4, the means for Set 3 are larger than Set 4 and are similar to the means and standard deviations for the forty common core items in Table 3.

Results Of the Correlational Analyses for the Mantel-Haenszel Indices

Correlational Analyses for the White-White Comparisons

The Pearson product-moment correlations among the MH D-DIF values (below the diagonal) and the Spearman rank-order correlations for the MHCHIX (above the diagonal) for all white-white comparisons are presented in Table 5.

Insert Table 5 here

The correlations among these indices would be expected to be low confirming a lack of differential item functioning since none would be expected to exist conceptually (Shepard, Camilli and Williams, 1984). The correlations for the MH D-DIF among the four small core samples are, for the most part, low (r=.-.13 to .26). The correlations for the small white core samples with the large white core sample are larger. Shepard et al.(1984) suggest that with correlations from overlapping samples, as is the case for the small core samples with the large core samples, consistent sampling error would be present. As there were few items flagged (two across all five analyses), these correlations probably reflect what Shepard et al.(1984) called spurious "differential item functioning" from common sample characteristics. The Spearman rank-order correlations among the MHCHIX statistics shown in Table 5 are similar to the correlations among the MH D-DIF estimates indices.

Correlations Among MH D-DIFF Values and MHCHIX
Forty Common Core Items: Black-White Comparisons

Correlations among
Overlapping Samples

The correlations for the four core samples (s1, s2, s3, s4) with the large sample (S) for the MH D-DIF values (below the diagonal) and MHCHIX (above the diagonal) are presented in Table 6. The correlations for the MH D-DIF calculated with Criterion 1 (total score on the forty common core items) range from .74 to .88. while the correlations for the MH D-DIF calculated with Criterion 2 (total score on 75 item test under study) are between .75 to .88 suggesting that the MH D-DIF is robust to item context effect. The correlations for the MHCHIX under Criterion 1 and Criterion 2 are lower than the corresponding MH D-DIF correlations. When looking at the correlations for Criterion 1 in contrast to Criterion 2 only the s1 correlations are different.

Correlations among MH D-DIF Values

Table 7 presents a multi-criterion, multi-sample matrix (MCMS). Conceptually, this matrix can be considered analogous to a multi-trait, multi-method matrix. The Pearson product-moment correlations for MH D-DIF values are below the main diagonal while the Spearman rank-order correlations for the MHCHIX values are above the main diagonal.

Insert Table 7 here

The correlations among the small core samples under Criterion 1 in the upper left-hand corner essentially

indicate the agreement among the MH D-DIF for the same items given to four different groups of examinees. These coefficients range from .36 to .60. These are low suggesting these indices are sample specific at least with samples of this size. The correlation between s1 and s4 is the lowest. A plot of the MH D-DIF values for s1 and s4 was inspected. For several items, the MH D-DIF values were considerably different. For instance, on item 2, the MH D-DIF value in s1 was .14. In contrast, in s4, the MH D-DIF was 1.16.

The correlations among the MH D-DIF values for the core items calculated with Criterion 2 (lower right hand corner) are also low within the different test taker samples. Again the lowest correlation is between s1 and s4. The MH D-DIF values were again plotted and showed much the same effect (For item 26, 1.9 vs. -.13). Comparing the correlations for Criterion 1 (upper left hand corner) and Criterion 2 (lower right-hand corner), there is little variation.

The correlations for the MH D-DIF values calculated with Criterion 1 and the MH D-DIF values associated with Criterion 2 among the same and different samples of test takers are presented in the lower left-hand corner of Table 7. The correlations between the MH D-DIF values calculated with Criterion 1 and the MH D-DIF values with Criterion 2 for the same samples of test takers shown in the diagonal are in the .90s again suggesting minimal context effects. In contrast, the correlations in the off-diagonals (for the MH D-DIF values from Criterion 1 and the MH D-DIF values

computed with Criterion 2) among different samples of test takers are lower (.35-.61). For instance, the correlation between the MH D-DIF values from s1 under Criterion 1 and the MH D-DIF values from s2 under Criterion 2 is .55. Again the correlations between the s1 and s4 samples are lowest reflecting the large differences in MH values from these samples for specific items.

Correlations among MHCHIX

The Spearman rank-order correlations above the diagonal in the upper left-hand corner indicate the variation in the MHCHIX values calculated with the same criterion (Criterion 1) for the same items among four samples of test takers (s1, s2, s3, s4). They range from .14 to .36. These are considerably lower than the corresponding correlations for the MH D-DIF values (below the diagonal in the upper left hand corner). The correlations for the MHCHIX calculated under Criterion 2 are in the lower right-hand corner above the diagonal. These are lower than the correlations for Criterion 1 suggesting the MHCHIX is sensitive to item context effects.

The upper right-hand corner lists the correlations for the MHCHIX calculated with Criterion 1 and the MHCHIX from Criterion 2. The correlations between the same sample of test takers under the different criterion are contained in the diagonal and range from .71 to .91. The correlation for s1 (.71) is considerably lower than the others. A plot of the MHCHIX values from s1, Criterion 1 and s1, Criterion 2

was inspected. There were differences in the MHCHIX values for several items from the Criterion 1 analysis in comparison to the Criterion 2 analysis. The off-diagonals list the correlations from different samples of test takers for the MHCHIX computed with Criterion 1 and with Criterion 2. These correlations range from .09 to .32. These correlations are considerably lower than those for the corresponding MH D-DIF values.

Correlations for MH-DIFF and MHCHIX: Rotated Test Forms

The correlations for the matching rotated test forms from the MH D-DIF values and MHCHIX calculated with Criterion 2 and those calculated with Criterion 3 (total score on 35 item rotated test form under study) were calculated. The Pearson product-moment correlations for the MH D-DIF values, were similar, as expected to the corresponding correlations found in the diagonals of Table 7. They ranged from .97 for s1 to .99 for s4. The Spearman rank-order correlations calculated for the MHCHIX values ranged from .63 for s1 to .87 for s4 and are similar in magnitude to analogous correlations from the forty common core items.

Standard Errors of MH D-DIF Indices

The mean standard error of the MH D-DIF indices for the total sample black-white comparison for the 40 item core test was .011. For the indices from the smaller samples, the means ranged from .044 to .052. Since the large core sample (S) is roughly four times larger than the small core samples

(s1, s2, s3, s4), a four to one ratio of standard errors is the order of magnitude difference that would be expected. The standard errors for the analogous white-white comparisons were similar.

DISCUSSION

Stability of the Mantel-Haenszel Estimates

The correlational analyses of the Mantel-Haenszel indices suggest that with samples of test takers of the sizes used for most of this study, the MH procedure is susceptible to idiosyncratic sample characteristics. The correlations between different samples of test takers taking the same items were low suggesting that larger sample sizes than those used for the majority of the analyses in this investigation are necessary to obtain stable estimates from the Mantel-Haenszel procedure. The standard errors for the MH D-DIF indices calculated for the smaller samples are larger than those from the large core sample. This size of the standard errors were similar for both the black- white comparisons and the white-white comparisons.

Item Context Effects

The correlational analyses suggest that the MH D-DIF is robust to item context effects. The correlations for the small core samples from Criterion 1 and Criterion 2 were similar in range and size.

Differential Item Functioning

Four items were identified as functioning differently for black and white test takers from analysis of the Core

items using the total sample. The two items favoring white test takers were classified at the application level. One was a measurement item with a picture of ruler asking for an estimate of where the ruler was positioned. The other was an algebra story problem with a simple text. Two items classified at the computation level favored black test takers. Both were simple mathematical sentences, one involving signed arithmetic, the other fractions. The same trend appeared in the analyses with the smaller sample sizes. In general, items classified at the application level, which involved reading as a first step in the solutions favored white test takers, while algebra and arithmetic items classified at the computation level with minimal reading favored black test takers. However, this should be interpreted with caution because of the instability of the estimates from the smaller samples of test takers.

When looking at patterns of identification across all analyses, what is interesting is the fact that more items from the small samples of test takers are flagged than the analysis using the large sample. Items which could have been flagged up to nine times were usually identified only twice. However no items identified as favoring one group in one analysis favored the other group in another. The direction remained the same or the item functioned the same for both groups.

Zwick and Ericikan (1989) found that controlling on additional variables shifted values for MHCHIX for all comparisons. The values of both MH D-DIF and MHCHIX shifted for the Hispanic-white comparisons, in particular. The sample sizes for the Hispanic-white comparisons in Zwick and Ericikan (1989) were only slightly larger than the samples sizes for the black- white comparisons in this study. Wild (1987) suggests that very large sample sizes are necessary to obtain the replicability of differential item functioning across different samples of test takers.

The Mantel-Haenszel procedure is a useful addition to the detection of differential item functioning. Additional investigations as well as simulation studies systematically varying the sample size of test takers to determine what sample size is needed to obtain stable estimates would be of interest.

REFERENCES

Angloff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Bloom, B. S. (Ed.), (1956). Taxonomy of educational objectives: The classification of educational goals. (Handbook 1). New York: McKay.

Camilli, G. (1979). A criticism of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

Cole, N. S. (1981). Bias in testing. American Psychologist, 36, 1067-1077.

Control Data Systems, (1985). Fortran Version 5 Reference Manual. Sunnyvale, CA: Control Data Corporation.

Green, D. R., & Draper, J. F. (1972). Exploratory studies of bias in achievement tests. Paper presented at the annual meeting of the American Psychological Association, Honolulu, Hawaii. (ERIC document Reproduction Service No. Ed 070 794).

Holland, P. W. (1985). On the study of differential item difficulty without IRT. Proceedings of the Military Testing Association, in press.

Holland, P. W. & Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. (Research Report No. 86-31). Princeton: Educational Testing Service.

Linn, R. L., & Drasgow, F. (1987). Implications of the golden rulesettlement for test construction. Educational Measurement: Issues and Practices, 6(2), 13-17.

Linn, R. L., Levin, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Lord, F. M. (1977). Practical applications of item response characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748

McPeek W. M., & Wild C. L. (1986). Performance of the Mantel-Haenszel statistic in a variety of situations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Phillips, A. & Holland, P. W. (1987). Estimators of the variance of the Mantel-haenszel loggs-odds-ratio estimate. Biometrics, 43, 425-431.

Scheuneman, J. D. (1979). A new method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Second International Mathematics Study. (1985). Summary report for the United States. Champaign, IL: Stipes.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Wigdor, A. K. & Garner, W. R. (Eds.), (1982). Ability Testing: Uses, consequences, and controversies. Part I: Report of the committee on ability testing. Washington, DC: National Academy Press.

Wild, C. (1987, May). Summary of DIF findings from recent research and use. Panel to Review ETS Index of Differential Item Difficulty. Princeton: Educational Testing Services.

Zwick, R & Ercikan, K., (1989). Analysis of Differential Item Functioning in the NAEP History Assessment. Journal of Educational Measurement, 26, 1, 55-66.

Table 1
SIMS Mathematics Tests: Descriptive Statistics

Test Form	Number of items	Coefficient Alpha	Mean	S.D.	Mean p
Core	40	.91	19.77	9.06	.49
TF1S	35	.85	16.98	6.65	.49
TF2S	35	.85	15.39	6.62	.44
TF3S	35	.86	16.44	6.80	.47
TF4S	35	.85	17.24	6.85	.49
TF1L	75	.94	37.24	15.07	.50
TF2L	75	.94	35.12	14.98	.47
TF3L	75	.94	36.32	15.24	.48
TF4L	75	.94	36.95	14.96	.49

Table 2
Mantel Haenszel Analyses Design

Test Form	Code	N	Matching Criterion	Items Tested
		Set 1		
White-focal Core	CoreW	4879	Core Total	Items1-40
White-focal Test Form1	TF1W	1207	Core Total	Items1-40
White-focal Test Form2	TF2W	1186	Core Total	Items1-40
White-focal Test Form3	TF3W	1230	Core Total	Items1-40
White-focal Test Form4	TF4W	1205	Core Total	Items1-40
		Set 2		
Core	Core	5685	Core Total	Items1-40
Blacks		670		
Whites		5015		
Test Form1 Core	Core1	1405	Core Total	Items1-40
Blacks		143		
White		1262		
Test Form2 Core	Core2	1407	Core Total	Items1-40
Blacks		167		
White		1240		
Test Form3 Core	Core3	1400	Core Total	Items1-40
Blacks		179		
Whites		1221		
Test Form4 Core	Core4	1381	Core Total	Items1-40
Blacks		141		
Whites		1240		
		Set 3		
Test Form1 Long	TF1L	1407	Core+TF1 Total	Items1-75
Black		167		
White		1240		
Test Form2 Long	TF2L	1400	Core+TF2 Total	Items1-75
Black		179		
White		1221		
Test Form3 Long	TF3L	1432	Core+TF3 Total	Items1-75
Black		169		
White		1263		
Test Form4 Long	TF4L	1381	Core+TF4 Total	Items1-75
Black		141		
White		1240		
		Set 4		
Test Form1 Short	TF1S	1436	TF1 Total	Items1-35
Blacks		172		
Whites		1264		
Test Form2 Short	TF2S	1433	TF2 Total	Items1-35
Blacks		187		
Whites		1246		

-continued on the next page-

Table 2, continued

Test Form3 Short	TF3S	1452	TF3 Total	Items1-35
Black		173		
White		1279		
Test Form4 Short	TF4S	1405	TF4 Total	Items1-35
Black		143		
White		1262		

N=number of examinees

Table 3
Descriptive Statistics for MHCHIX and MH D-Dif
Sets 1, 2, and 3: Forty Common Core Items

Test Form	MHCHIX		MH D-DIF	
	Mean	S.D.	Mean	S.D.
CoreW	1.1	1.6	.005	.26
TF1W	1.1	1.5	.002	.56
TF2W	.75	.90	.006	.50
TF3W	1.2	2.1	.012	.56
TF4W	.77	1.3	.005	.48
		Set 2		
Core	6.6	8.9	-.017	.65
Core1	2.0	3.0	-.003	.75
Core2	3.3	4.6	-.008	.91
Core3	2.2	2.4	-.031	.83
Core4	1.8	2.5	-.036	.77
		Set 3*		
TF1L	2.2	3.3	.241	.75
TF2L	3.0	4.0	.123	.87
TF3L	2.1	2.4	.025	.81
TF4L	1.6	2.4	.091	.70

Note. Values are based on the 40 common core items only.

Table 4
Descriptive Statistics for MH-D Dif and MHCHIX:
Sets 3, and 4: Rotated Test Forms

Test Form	MHCHIX		MH D-DIF	
	Mean	S.D.	Mean	S.D.
TF1L	2.0	2.3	-.231	.70
TF1S	1.8	2.6	.039	.69
TF2L	2.8	4.0	-.177	.83
TF2S	3.2	4.2	-.017	.90
TF3L	1.4	2.0	-.038	.65
TF3S	1.7	2.4	.028	.70
TF4L	1.6	2.1	-.137	.68
TF4S	1.4	2.0	-.012	.68

Note. All "L" statistics are based on last 35 items of the 75 item test

Table 5
Correlations among the MH D-DIF and MHCHIX
for the White-white Comparisons: Set 1

	COREW	TF1W	TF2W	TF3W	TF4W
COREW	*	.13	.24	.44	.26
TF1W	.57	*	.32	.18	-.01
TF2W	.33	-.13	*	-.14	-.10
TF3W	.68	.26	-.10	*	-.02
TF4W	.44	-.11	.11	.12	*

Note. All correlations are based on the 40 core items.

Table 6
Correlations Matrix
for the Core Test Items: MH D-DIF

		Criterion 1				Criterion 2			
	Core	Core1 (s1)	Core2 (s2)	Core3 (s3)	Core4 (s4)	TF1L (s1)	TF2L (s2)	TF3L (s3)	TF4L (s4)
Core	*	.61	.66	.42	.45	.38	.63	.45	.39
(s1)	.77								
(s2)	.88								
(s3)	.83								
(s4)	.74								
(s1)	.75								
(s2)	.88								
(s3)	.82								
(s4)	.76								

Note. All correlations based on indices from 40 core items.

Table 7
Multi-criterion Multi-sample Correlation Matrix
for the Core Test Items: MH D-DIF and MHCHIX

	Criterion 1				Criterion 2			
	Core1 (s1)	Core2 (s2)	Core3 (s3)	Core4 (s4)	TF1L (s1)	TF2L (s2)	TF3L (s3)	TF4L (s4)
(s1)	*	.36	.32	.20	.71	.11	.14	.09
(s2)	.56	*	.14	.26	.32	.90	.15	.26
(s3)	.58	.58	*	.15	.29	.26	.91	.17
(s4)	.36	.60	.52	*	.16	.25	.18	.91
(s1)	.98	.55	.56	.35	*	.17	.12	.06
(s2)	.57	.99	.61	.58	.59	*	.29	.27
(s3)	.57	.60	.98	.49	.56	.63	*	.22
(s4)	.41	.61	.54	.98	.39	.60	.50	*

Note. Correlations based on indices from 40 core items.